CLASSIFYING SOUND SCENES

Machine Learning for Cities
Spring 2016

Adhlere Coffy
Neil D'Souza
Maria Filippelli
Priya Khokher
Manushi Majumdar

Cities are filled with an abundance of sensory perceptions and, at times, the cacophony of sounds can prove to be overwhelming. But can this myriad of noises be utilized to identify a particular geographical area or “urban scene”? Can cities be distinguished from each other based on the sound scenes that are prevalent within them? This undertaking sought to execute preliminary machine learning analyses to investigate the question: Are urban sound profiles from different cities similar enough to accurately classify, regardless of origin?

By utilizing a variety of machine learning techniques, this project classified common city sound scenes in an effort to determine if urban sound scenes are generalizable and which model achieved the greatest accuracy. Sound scenes were taken from datasets developed in London and Paris and transformed using Mel-frequency cepstral coefficients (MFCCs) in order to create values to test with Naive Bayes, Support Vector Machines, and Neural Networking models. These models were chosen because they are the most applicable to sound scene classification since they are better at handling the kind of features that are involved with sound (i.e. the specific numeric MFCC values).

London and Paris have attempted similar classifications and provided open datasets for this project. Five urban scenes common to both datasets were used for classification, a subway station, restaurant, busy street, farmer’s/open air market, and a quiet street. These scenes were selected for the wide range of urban activities that they represented, are distinctly different from each other, and presence in both datasets. An additional group was reserved for “general” scenes that were not cross compatible between datasets.

Outside of the London and Paris research, much of the existing sound classification has focused on classifying sound sources, not scenes. Sound sources merely focus on extracting the information on the source of a particular sound, while this undertaking endeavored to classify sound scenes without separating sound sources. Results showed that the classification of sound scenes could open doors to deeper research in this field.

Datasets

The two datasets used are from London and Paris sound scene classifications. The Detection and Classification of Acoustic Scenes and Events (DCASE) Queen Mary dataset, an urban sound scene classification experiment done in London is comprised of 10 scene classes with 20 sound clips each. The Rouen dataset is from an urban sound scene classification experiment in Paris and has 19 sound scenes. Several sound scenes exist in both datasets making them ideal to train, validate, and test with our models. The scenes common to both and chosen for classification are restaurant, busy street, open air/farmer's market, and subway station, any additional scenes were classified as other. The sound clips from both datasets were 30 seconds each which is an ideal minimum length for testing.

Data Wrangling

Classification of urban sound scenes entailed categorization of data points based on their predefined set of values. The sounds scenes were transformed into a feature space of frequency and amplitude and required several transformations to prepare the data for model testing. The creation of a feature space for the sound dataset using frequency instead of amplitude as frequencies created a more specific and inherent to a sound type than amplitude alone because it externally manipulated. A new set of coefficients - Mel Frequency Cepstral Coefficients (MFCCs) have proven to form a more robust feature space and have been used for the project. As sound datasets can contain redundancies in information, it was necessary to choose a feature set as invariant to changes within the natural class' range as possible.

The next step was to produce MFCCs for feature extraction. MFCCs were calculated in three steps, first a fourier transformation was performed, then powers were mapped onto a Mel scale, and finally the MFCC of each spectrum was produced as shown in the following diagram.

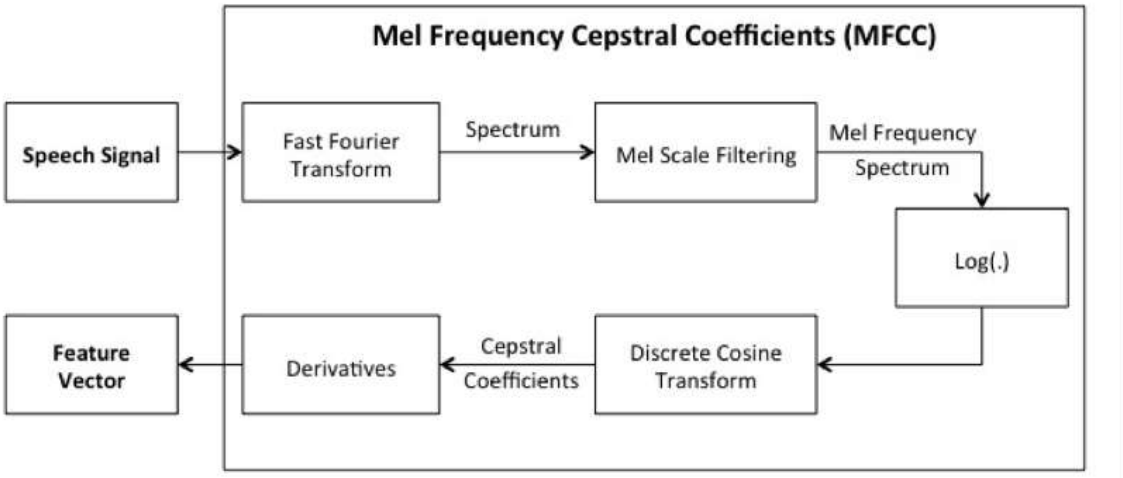

Image credit: *http://tinyurl.com/hjt34un*

The fourier transformation divided each sound scene into 20 millisecond frames. This was important because sound scenes differ in the patterns they produce and no reasonable computation can be performed by taking the mean of longer frames. The fourier transform was then used to compute the frequency spectrum of the frame.

Mel scaling related the perceived frequency of a pure tone to its actual measured frequency. Incorporating this scale allowed the features match more closely what humans hear. To achieve this task, 40 different mel filter banks were used to bin each frequency change, ranging from fmin = 0 to fmax = 22050 Hz.

To find the resulting MFCCs, the log of each Mel power frequency was taken and then the discrete cosine transformation. For each set of MFCCs, each frame included (there were 40 and the last 20 were used as these capture less noise) a summarization of : minimum, maximum, median, mean, variance, and kurtosis. From this the mean and variance of the first derivatives were chosen as features because this is general practice.

Methods:

After extracting the MFCC and calculating appropriate summary statistics (e.g. mean and variance), the data was separated twice for two separate sets of testing. In the first separation, the datasets were combined and then randomly split into training, test, and

validation sets. This allowed for evaluation of whether or not the hypothesis, that urban sound scenes are common to all urban environments, could be adapted for sound and used as an additional approach to understanding urban environments.

In the second separation, the data was randomly shuffled such that the training, validation, and test sets reflected all of the available data. This was achieved by first concatenating the datasets and then randomly splitting the resulting dataset into a 60-20-20 ratio of training, validation, and testing data. This approach served as a means of determining the generalizability of the models.

The models chosen for evaluation were naïve bayes, neural networks, and support vector machines. The model assumptions were feature independence and supervised learning because all data is labeled.

Naive Bayes

Naive Bayes (NB) was performed first because of its streamlined implementation and classification methodology. For this model a continuous NB classifier was used to account for the number of scenes identified for classification and the corresponding feature set which was quite expansive. A primary consequence of note when using NB is that it assumes that features are independent of one another. This failure to identify correlations between features lends NB to underperform compared to other machine learning techniques as demonstrated by this project.

Using the aforementioned dataset split, the accuracy of the classifier between the training and validation splits were compared. The first dataset train/validation accuracy was an underwhelming 68%, a less than stellar performance. The second dataset yielded an accuracy of 22% thus proving that Naive Bayes was not a good classifier for the urban sound scenes.

Neural Networks

Neural Networks (NN) provide an intuitive, extremely flexible option for classifying data and making it an actionable all in one process. The difficulty that arises from the use of NN's is the time consuming nature of optimizing the performance. NN's are very easy to implement and

use but quite difficult to improve for the sake of finding the “best” outcome. Adjusting parameters such as learning rate, number of hidden layers, and batch size can be very time consuming without yielding results that demonstrate improvements in performance over time. Being based on the logistic regression, NN inherently utilizes probabilistic relationships between inputs, unlike NB, and does not assume that all inputs are independent of each other. A minor adjustment to the negative log likelihood function of utilizing the mean as opposed to the sum allows for reduction in dependence of the learning rate on the batch size. For this application, it was of interest to identify how well the NN model could learn from given sound scenes with a wide range of input magnitudes without severely impacting computational time and/or resources. Finale parameters for the model were as follows: Learning Rate - 0.85, Batch Size - 400, and 10 hidden layers.

Support Vector Machines

Support Vector Machines (SVM) utilized a technique called the kernel trick to transform the data and then, based on those transformations, computed an optimal boundary between the possible outputs discovered. SVM is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data, the algorithm outputs an optimal hyperplane which categorizes new examples.

A soft margin was used because it allowed for errors to be made while fitting the model to the training data set. Conversely, hard margins would have resulted in fitting of a model that allowed zero errors. It can be helpful to allow for errors in the training set, because it may produce a more generalizable model when applied to new datasets. Forcing rigid margins could have resulted in a model that performed perfectly in the training set, but over-fit or became less generalizable when applied to a new dataset. Soft margins acted as a penalty parameter.

Three types of kernels were tested, the linear, rbf or Gaussian, and polynomial. The polynomial kernel SVM models proved most successful with a higher accuracy than the other kernels. The model began with polynomial degree of two, making it quadratic. Changing the

degree from 2 to 3, increased the accuracy of the prediction 10%. Selecting a degree value of 3 allowed for a more flexible decision boundary without overfitting the data. The best accuracy was observed for penalty=2.5 with the training data is split approximately in 75:25 ratio.

Results:

MODELS (Best Split)	ACCURACY	
	Dataset 1	Dataset 2: Generalizability Set
Gaussian Naive Bayes (50:50)	68.4 %	21.8 %
Neural Networks (70:30)	78 %	80 %
Polynomial Support Vector Machine (~75:25)	93.7 %	91. 4%

The model that performs best on the validation set in both cases is SVM with a degree polynomial kernel and a penalty factor of 2.5. This is the model that is used on the test set as shown below.

TESTING	37.5 %	71.2%

Thus we can say that sound may not be the best way to demonstrate science of cities though it still shows some support. The accuracy over the generalized data is also very promising.

Conclusions and Future Work:

The SONYC project at CUSP has already delved into identifying sound sources in NYC, the next step would be to understand and identify the sound scenes which will provide the context for different sources. The classification will provide a basis for further research in the field of urban acoustics/sounds, which is still being explored. The primary take away, and relative success, of the project execution is that there are implications made that cities are in fact relatable in terms of typical sound scenes. The number of assumptions made here can be easily addressed with minimal effort made on the part of the experimenters, those being; increases to sample size and rate, additional sound profiles datasets from comparably sized cities, greater duration of time spent refining the models to ensure optimization has been met, and perhaps more extensive actions taken to randomize the information such that the degree of generalizability can be maximized.

RESOURCES

Bibliography:

Salamon, Justin, Christopher Jacoby, and Juan Pablo Bello. "A dataset and taxonomy for urban sound research." *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014.

Salamon, Justin, and Juan Pablo Bello. "Feature learning with deep scattering for urban sound analysis." *Signal Processing Conference (EUSIPCO), 2015 23rd European*. IEEE, 2015.

Salamon, Justin, and Juan Pablo Bello. "Unsupervised feature learning for urban sound classification." *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015.

Datasets:

A. Rakotomamonjy, G. Gasso, Histogram of gradients of Time-Frequency representations for audio scene detection, Technical report, HAL, 2014

D. Giannoulis, E. Benetos, D. Stowell, and M. D. Plumbley, "IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events - Public Dataset for Scene Classification Task", Queen Mary University of London, 2012.

Individual Contributions:

MFCC extraction and data wrangling - Priya

Gaussian Naive Bayes - Maria

Neural Networks - Adhlere

Support Vector Machine - Manushi

Testing and documentation - Neil

Code:

https://github.com/neilverosh/UASSC